statOmics

adriaan.sticker@ugent.be

# Robust summarization and inference in label-free quantification

Adriaan Sticker[1,2,3,4], Ludger Goeminne[1,2,3,4], Lennart Martens[2,3,4] & Lieven Clement[1,4]

[1] Department of Applied Mathematics, Computer Science & Statistics, Ghent University, Belgium [2] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
[3] UGent Department of Biomolecular Medicine, Ghent University, Ghent, Belgium [4] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium

## Background

Label-Free Quantitative (LFQ) mass spectrometry is standard for differential expression (DE) analysis, but protein abundance estimation remains challenging:

1. **Peptide-specific effects**: Different peptides from a protein have distinct physio-chemical properties leading to high variability in MS1 intensities.
2. **Context-sensitive missingness**: Peptides sometimes stay unidentified, especially low abundant ones co-eluting with high abundant ones get missing.

Summarization methods aggregate MS1 peptide to protein intensities and DE analysis is done on these protein summaries. The state-of-the-art peptide-based model MSqRob [1] tests for DE directly from peptide intensities using a linear model:

$$y_{itsp} = \beta_i^0 + \beta_{it}^{treatment} + \beta_{is}^{sample} + \beta_{ip}^{peptide} + \epsilon_i$$

With $y_{itsp}$ the normalised $\log_2$ intensity of peptide $p$ in protein $i$ from sample $s$ with treatment $t$. Parameter estimators variability is reduced through ridge penalisation, outliers are weighed down by M-estimation using Huber weights and the variance of $\epsilon_i$ is estimated with an empirical Bayes estimator. This protects against overfitting and gives more stable and accurate estimators, especially with few observations.

## Problem

The summarization strategies show huge differences in performance depending on the absolute and relative abundances of proteins between conditions and none clearly outperforms others. By reducing bias and improved uncertainty estimation, MSqRob has more precise and more accurate FC estimates then any other method and always outperforms the others for DE analysis (Figure 1).
MSqRob still has some drawbacks:

1. Fitting models on complex experimental designs is computational costly.
2. The residual degrees of freedoms are unclear due to nonrandom missingness.
3. Peptide random effects in a MSqRob mixed model often confuses end-users.
4. Protein summaries for visualisation or downstream processing are unavailable.

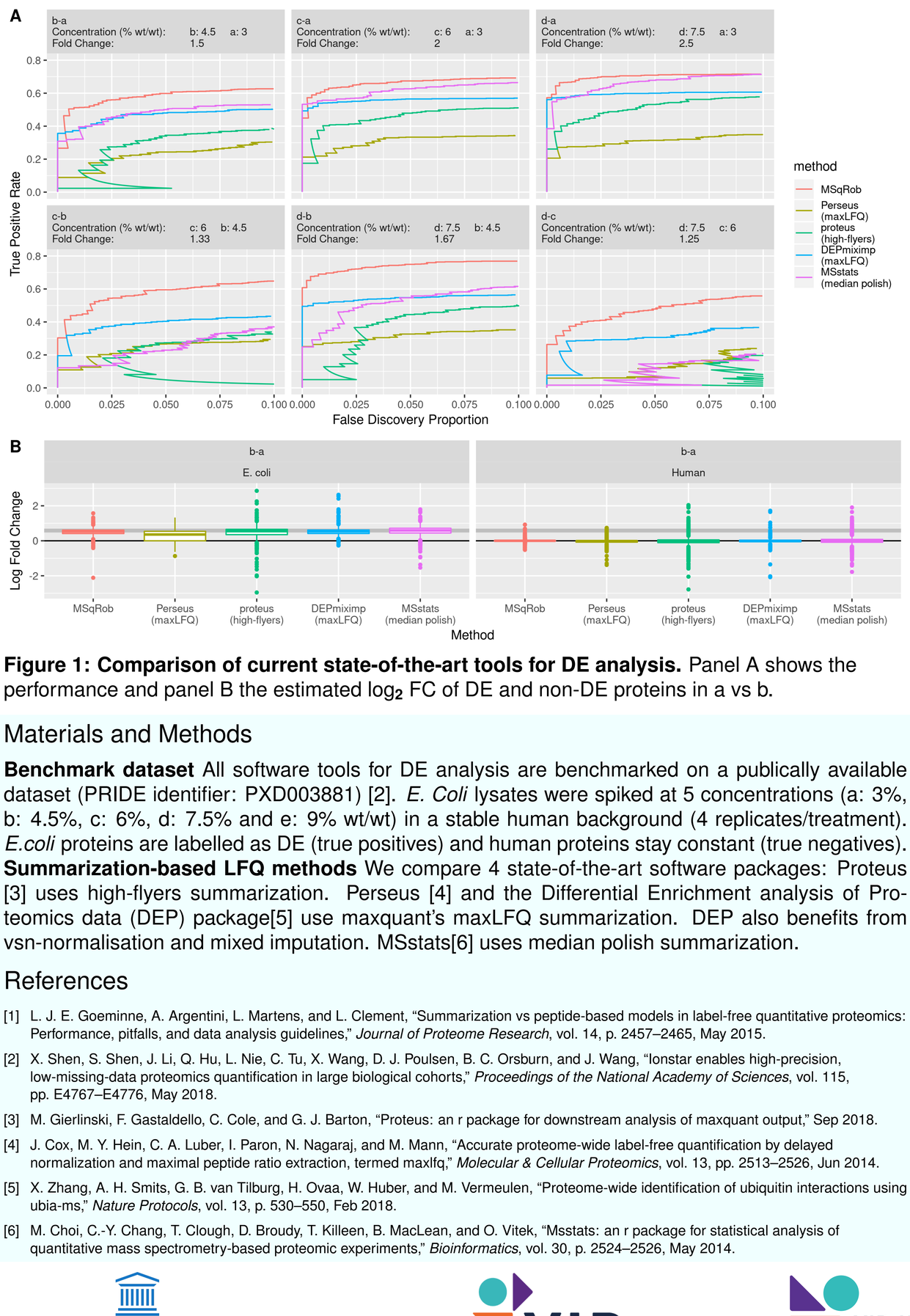


**Figure 1: Comparison of current state-of-the-art tools for DE analysis.** Panel A shows the performance and panel B the estimated $\log_2$ FC of DE and non-DE proteins in a vs b.

## Materials and Methods

**Benchmark dataset** All software tools for DE analysis are benchmarked on a publically available dataset (PRIDE identifier: PXD003881) [2]. *E. Coli* lysates were spiked at 5 concentrations (a: 3%, b: 4.5%, c: 6%, d: 7.5% and e: 9% wt/wt) in a stable human background (4 replicates/treatment). *E.coli* proteins are labelled as DE (true positives) and human proteins stay constant (true negatives).
**Summarization-based LFQ methods** We compare 4 state-of-the-art software packages: Proteus [3] uses high-flyers summarization. Perseus [4] and the Differential Enrichment analysis of Proteomics data (DEP) package[5] use maxquant's maxLFQ summarization. DEP also benefits from vsn-normalisation and mixed imputation. MSstats[6] uses median polish summarization.

## References

[1] L. J. E. Goeminne, A. Argentini, L. Martens, and L. Clement, "Summarization vs peptide-based models in label-free quantitative proteomics: Performance, pitfalls, and data analysis guidelines," *Journal of Proteome Research*, vol. 14, p. 2457–2465, May 2015.

[2] X. Shen, S. Shen, J. Li, Q. Hu, L. Nie, C. Tu, X. Wang, D. J. Poulsen, B. C. Orsburn, and J. Wang, "Ionstar enables high-precision, low-missing-data proteomics quantification in large biological cohorts," *Proceedings of the National Academy of Sciences*, vol. 115, pp. E4767–E4776, May 2018.

[3] M. Gierlinski, F. Gastaldello, C. Cole, and G. J. Barton, "Proteus: an r package for downstream analysis of maxquant output," Sep 2018.

[4] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann, "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq," *Molecular & Cellular Proteomics*, vol. 13, pp. 2513–2526, Jun 2014.

[5] X. Zhang, A. H. Smits, G. B. van Tilburg, H. Ovaa, W. Huber, and M. Vermeulen, "Proteome-wide identification of ubiquitin interactions using ubia-ms," *Nature Protocols*, vol. 13, p. 530–550, Feb 2018.

[6] M. Choi, C.-Y. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean, and O. Vitek, "Msstats: an r package for statistical analysis of quantitative mass spectrometry-based proteomic experiments," *Bioinformatics*, vol. 30, p. 2524–2526, May 2014.

## Solution

We propose a novel summarization strategy, **MSqRobSum**, which trains MSqRob in a two-stage procedure circumventing the drawbacks of MSqRob.

1. Peptide intensities of a protein are aggregated using robust regression with M-estimation using Huber weights using a protein-wise linear model:

$$y_{isp} = \beta_{is}^{sample} + \beta_{ip}^{peptide} + \epsilon_i$$

   By encoding the peptide effect as a sum contrast, $\beta_{is}^{sample}$ can be interpreted as the mean intensity for protein $i$ in sample $s$.
2. A simplified MSqRob analysis is done on these protein summaries:

$$y_{it} = \beta_i^0 + \beta_{it}^{treatment} + \epsilon_i$$

MSqRobSum has similar performance to MSqRob for medium to highly DE proteins and only breaks down at increasingly lower fold changes in DE (Figure 2). However, all summarization methods drop in performance at lower fold changes. The FDR is always correctly controlled except in comparisons c-a and d-a due to ion competition.

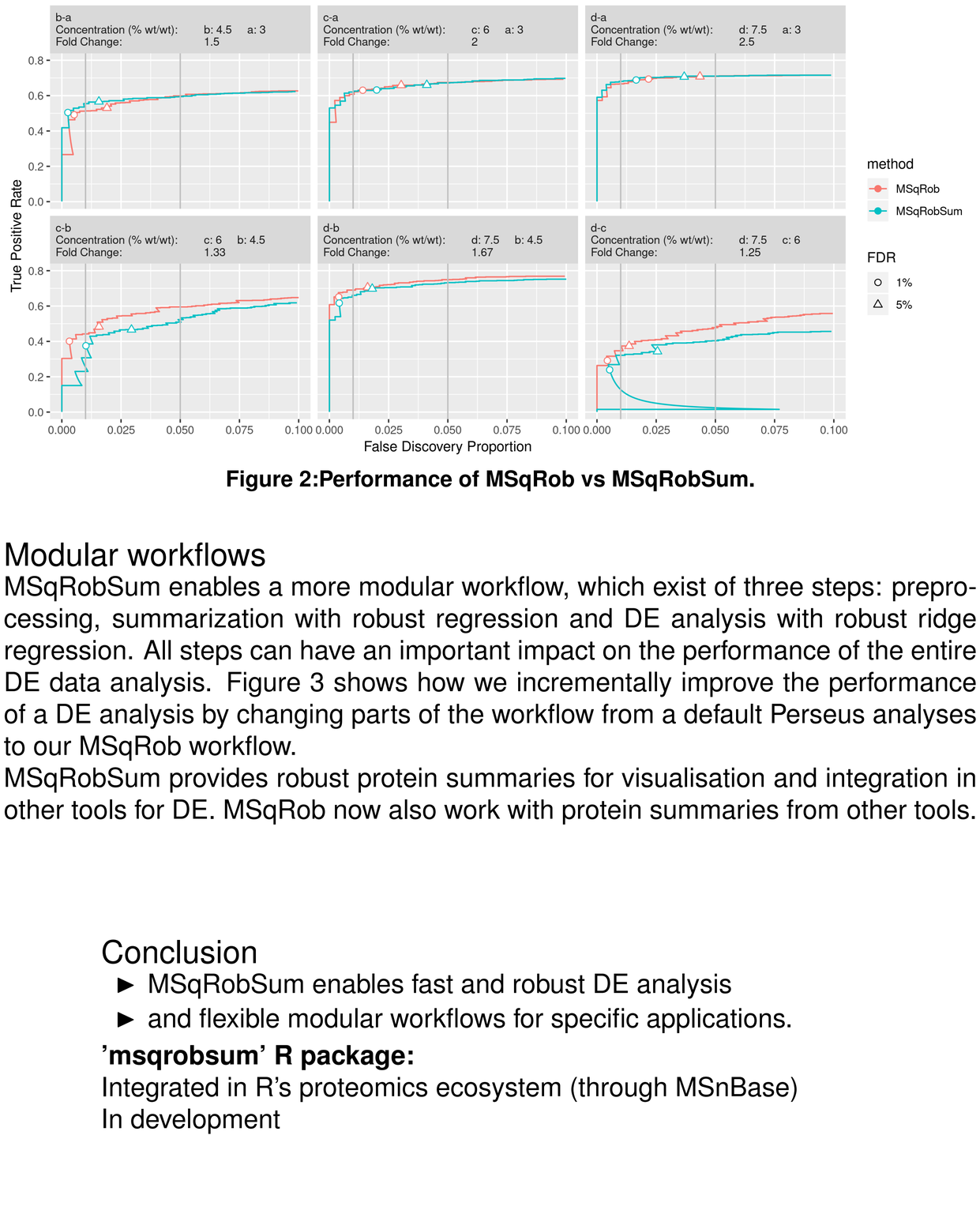


**Figure 2:Performance of MSqRob vs MSqRobSum.**

## Modular workflows

MSqRobSum enables a more modular workflow, which exist of three steps: preprocessing, summarization with robust regression and DE analysis with robust ridge regression. All steps can have an important impact on the performance of the entire DE data analysis. Figure 3 shows how we incrementally improve the performance of a DE analysis by changing parts of the workflow from a default Perseus analyses to our MSqRob workflow.
MSqRobSum provides robust protein summaries for visualisation and integration in other tools for DE. MSqRob now also work with protein summaries from other tools.

## Conclusion

- MSqRobSum enables fast and robust DE analysis
- and flexible modular workflows for specific applications.

**'msqrobsum' R package:**
Integrated in R's proteomics ecosystem (through MSnBase)
In development

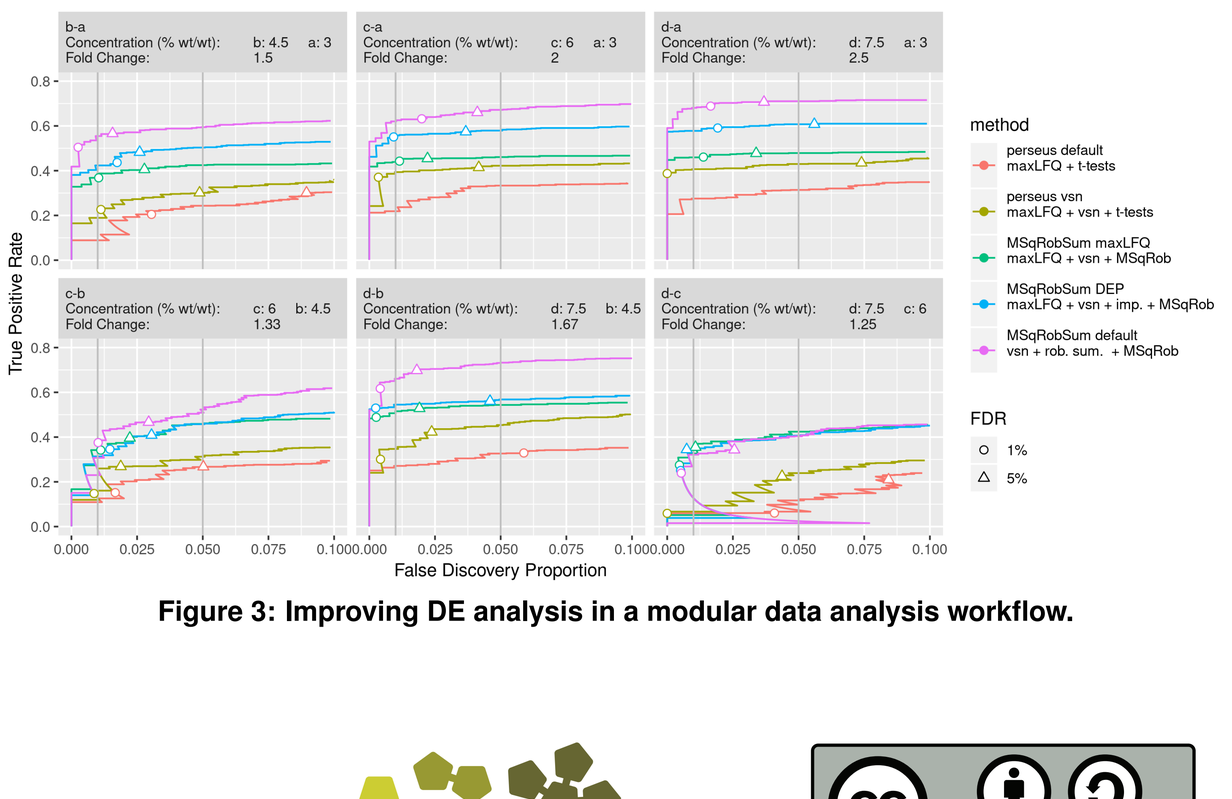


**Figure 3: Improving DE analysis in a modular data analysis workflow.**

GHENT UNIVERSITY

VIB science meets life

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

BIG N2N Bioinformatics Institute Ghent From nucleotides to networks